

FINAL PROJECT

Noam Atias 311394357
Chanel Michaeli 208491787

Kaggle Competition - Evaluating Student Writing

Goal:

On this competition we will identify elements in student writing.

More specifically, we will automatically segment texts and classify argumentative and rhetorical elements in essays written by 6th-12th grade students.

General Outline

The general outline of our representation is as follows -

- ❑ Data analysis- exploring the data.
- ❑ Defining the problem and the tools for solving it.
- ❑ Preprocessing - preparing the labels and the input data.
- ❑ Model Building and Training – Long Former model.
- ❑ Evaluation.
- ❑ Submission.

Data exploration

Training data consists of 15594 texts.

Training data consists of 144293 annotations.

Each essay contains average 9.3 annotations.

Discourse elements are:

- Lead
- Position
- Claim
- Counterclaim
- Rebuttal
- Evidence
- Concluding Statement

Texts Analysis:

Each text has an ID and each discourse element has an ID.

Discourse Start- Character position where discourse element begins in the essay response.

Discourse End - Character position where discourse element ends in the essay response

Discourse Text - Text of discourse element

Discourse Type - Classification of discourse element

Discourse Type Number - enumerated class label of discourse element

Prediction String - The word indices of the training sample, as required for predictions

	id	discourse_id	discourse_start	discourse_end	discourse_text	discourse_type	discourse_type_num	predictionstring
0	423A1CA112E2	1.622628e+12	8.0	229.0	Modern humans today are always on their phone....	Lead	Lead 1	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 1...
1	423A1CA112E2	1.622628e+12	230.0	312.0	They are some really bad consequences when stu...	Position	Position 1	45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
2	423A1CA112E2	1.622628e+12	313.0	401.0	Some certain areas in the United States ban ph...	Evidence	Evidence 1	60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
3	423A1CA112E2	1.622628e+12	402.0	758.0	When people have phones, they know about certa...	Evidence	Evidence 2	76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 9...
4	423A1CA112E2	1.622628e+12	759.0	886.0	Driving is one of the way how to get around. P...	Claim	Claim 1	139 140 141 142 143 144 145 146 147 148 149 15...

Example of text and its discourse elements:

Hi, i'm Isaac, i'm going to be writing about how this face on Mars is a natural landform or if there is life on Mars that made it. The story is about how NASA took a picture of Mars and a face was seen on the planet. NASA doesn't know if the landform was created by life on Mars, or if it is just a natural landform. ---> Lead

On my perspective, I think that the face is a natural landform because I don't think that there is any life on Mars. In these next few paragraphs, I'll be talking about how I think that it is a natural landform ---> Position

I think that the face is a natural landform because there is no life on Mars that we have discovered yet ---> Claim

If life was on Mars, we would know by now. The reason why I think it is a natural landform because, nobody lives on Mars in order to create the figure. It says in paragraph 9, "It's not easy to target Cydonia," in which he is saying that it's not easy to know if it is a natural landform at this point. In all that they're saying, it's probably a natural landform. ---> Evidence

People thought that the face was formed by aliens because they thought that there was life on Mars. ---> Counterclaim

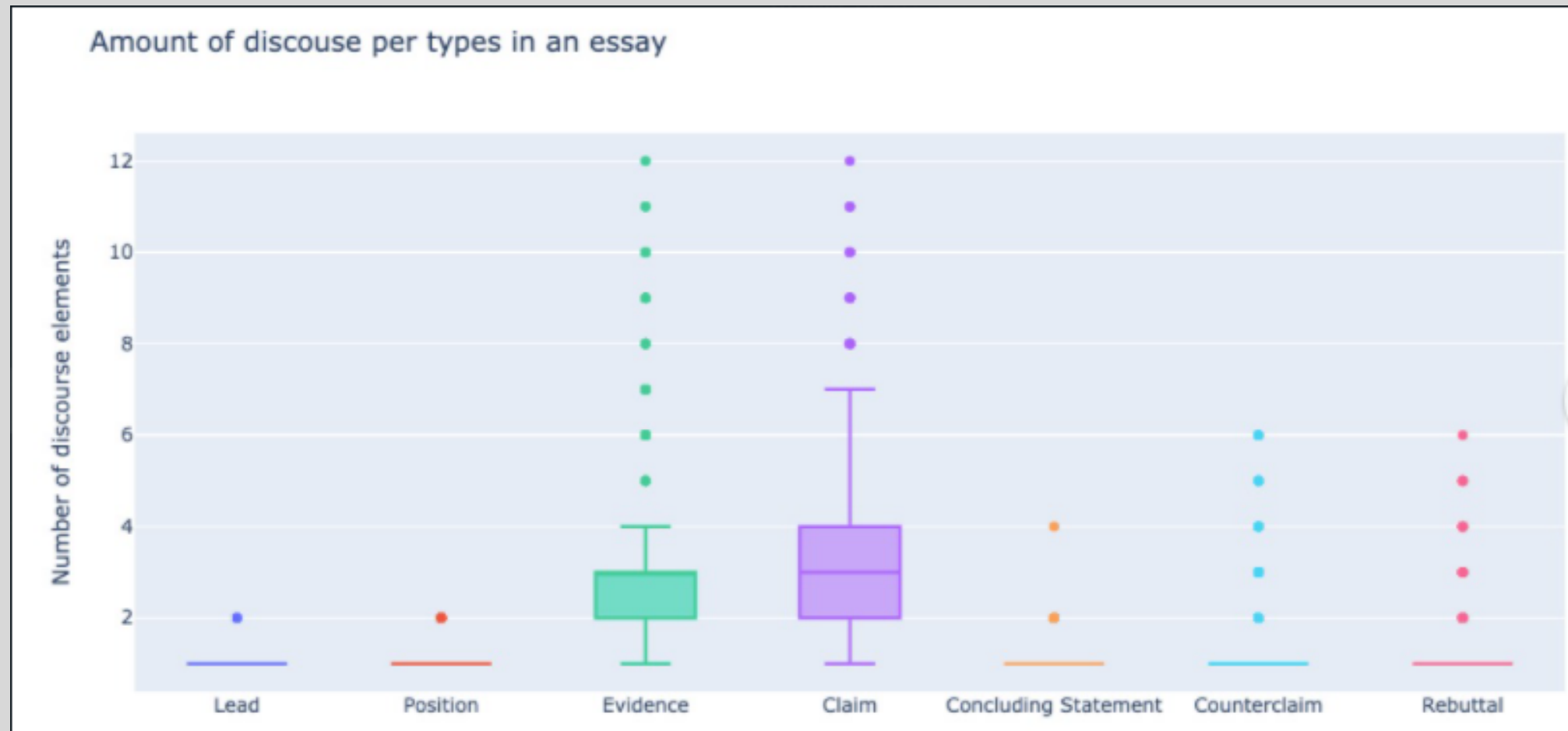
though some say that life on Mars does exist, I think that there is no life on Mars. ---> Rebuttal

It says in paragraph 7, on April 5, 1998, Mars Global Surveyor flew over Cydonia for the first time. Michael Malin took a picture of Mars with his Orbiter Camera, that the face was a natural landform. ---> Evidence

Everyone who thought it was made by aliens even though it wasn't, was not satisfied. I think they were not satisfied because they have thought since 1976 that it was really formed by aliens. ---> Counterclaim

Though people were not satisfied about how the landform was a natural landform, in all, we now know that aliens did not form the face. I would like to know how the landform was formed. we know now that life on Mars doesn't exist. ---> Concluding Statement

Let us check how many discourse types are usually written in an essay:



The discourse types Claim and Evidence have the highest median at 3 elements in an essay

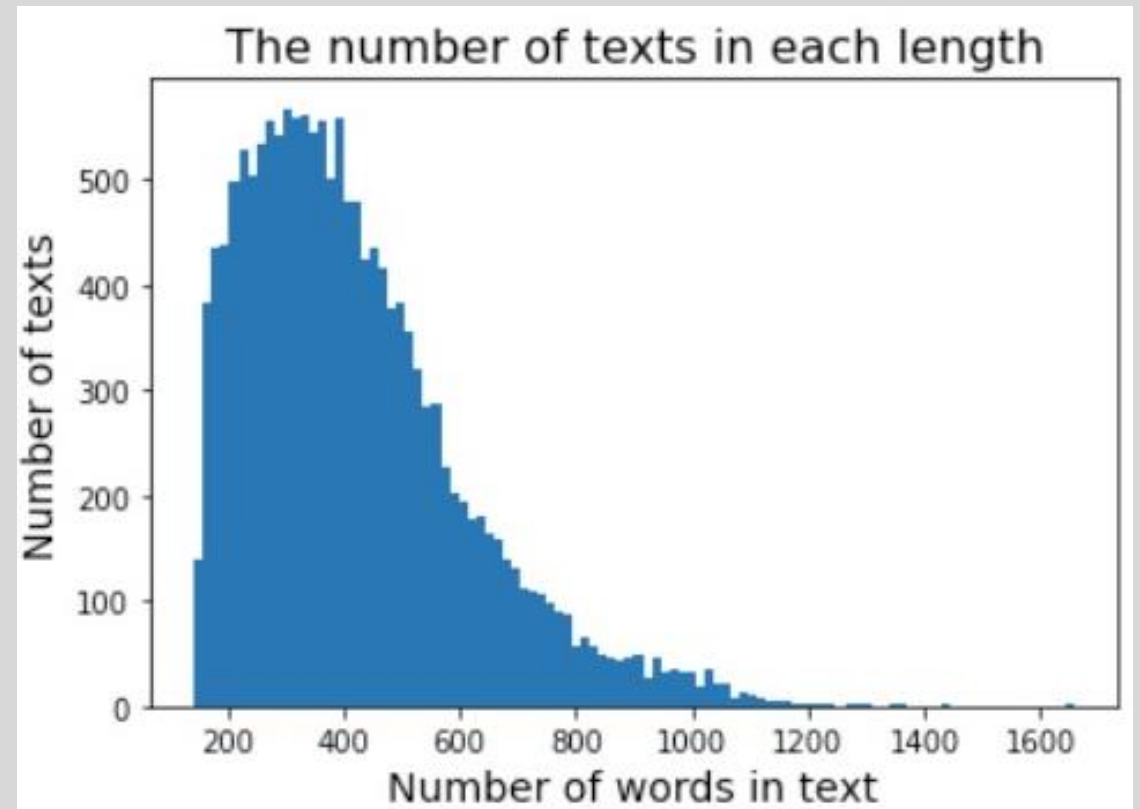
Maximum text length:

We can see that the number of words in most texts is smaller than 500 and most texts have less than 1000 words.

Therefore, using a transformer width of 1024 is a good comprise of capturing most of the data's signal but not having too large a model.

So, We determined the maximum length of each text to be 1024.

Texts that has less than the maximum length defined will be padded in the preprocessing stage and texts that has more than the maximum length will be truncated.



Defining the problem - NER

- NER, short for, Named Entity Recognition is a standard Natural Language Processing problem which deals with information extraction.
- The primary objective is to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, events, etc.
- In our case, we would like to locate and classify named entities in text which are the annotations into predefined categories which are the discourse elements.

Solving the problem - Transformers

Transformers is a library of state-of-the-art pre-trained models for Natural Language Processing (NLP).

Main structure blocks and their purpose:

1. The input sequence:

- The input sequence can be passed in parallel and determined the word embedding simultaneously.
- The idea is to map every word to a point (vector) in space, called embedding space, where similar words in meaning are physically closer to each other.
- The same word in different sentences may have different meaning. This is where positional encoders come in. It's a vector that has information on distances between words in a sentence.

2. Self-attention –

- One of the problems of recurrent models is that long-range dependencies (within a sequence or across several sequences) are often lost.
- Self attention works by considering or paying attention to all the other words, i.e., how relevant each word in a sequence to other words in the same sequence.
For every word we can have an attention vector generated which capture contextual relationships between words in a sequence.
- Therefore, self-attention solves the problem due to its ability to attend to different positions of the input sequence.
- Transformer creates stacks of self-attention layers that is called multi-head attention.
- The attention function:

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Long-former model

- Long-Former is a modified Transformer architecture that aimed for long documents.
- Traditional Transformer-based models are unable to process long sequences due to their self-attention operation, which scales quadratically with the sequence length.
- To address this, Long-Former uses an attention pattern that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer.
- Long-Former minimizes the computational burdens of attention by performing not all possible comparisons but only those with the relevant words.
- Long-Former provides several strategies to do this, such as letting most tokens only attend to their neighbors, while only a few attend to every token.
In other words, attention is now more local rather than global.
- We trained Long Former model with a maximum token length of 1024.

Preprocessing

1. Input Preprocessing:

- We need to convert the texts to sequences of token IDs, which will be used as indices into an embedding.
- We can use the functions provided by the transformers package to help us perform the tokenization and transformation easily. In particular, we can use the function `encode_plus`, which does the following in one go:
 - Tokenize the input sentence
 - Pad and truncate the sentence to the maximum length allowed.
 - Encode the tokens into their corresponding IDs Pad and truncate all sentences to the same length.
 - Create the attention masks which explicitly differentiate real tokens from [PAD] tokens.

2. Output Preprocessing:

- The targets of our task are the discourse types.
- We made a representation of the targets corresponding to the discourse elements in each text, i.e., for each word in each text, we represent its discourse type.
- We used a tagging format called Inside–outside–beginning tagging (IOB tagging) .
- The IOB format is a tagging format that gives us information about the location of the word in the chunk (in our case, the discourse elements).
- The IOB Tagging system contains tags of the form:
 - B– for the word in the Beginning chunk
 - I– for words Inside the chunk
 - O – Outside any chunk
- Therefore, we created 14 labels, 2 labels for each discourse type (I and B), and the last target is O if a word does not belong to any label.

Building the model

- We used Long-Former backbone
- We added our own NER head using:
 - one hidden layer of size 256
 - one final layer with SoftMax of size 15 (to output probabilities over 15 classes).
- We define the Adam optimizer, that is used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data.
- We defined the Cross-Entropy loss as categorical function loss since it's the most suitable loss for this kind of tasks.
- We split the texts to validation dataset and train dataset (to ensure that we are not overfitting).

Training the model

➤ **Choosing hyper-parameters:**

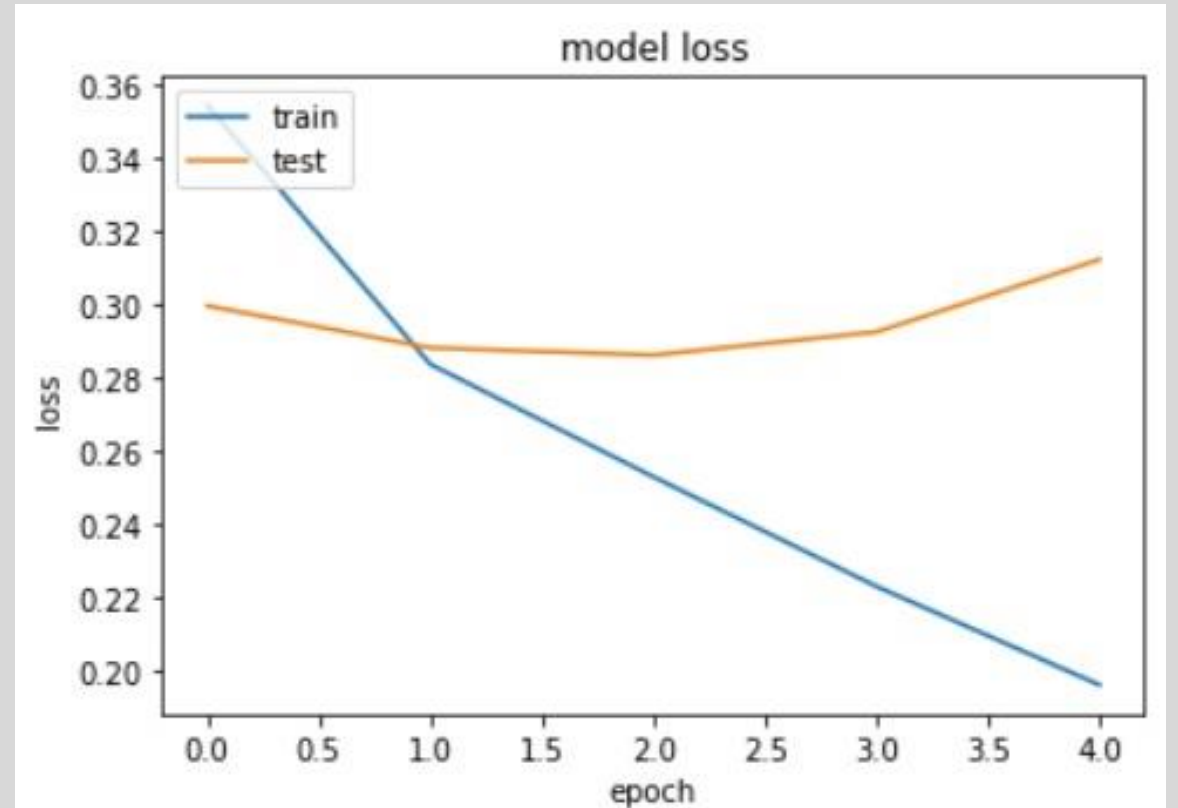
We trained the model several times to find the best hyper parameters:

- First, we took small learning rate, but we saw that the loss didn't change much from iteration to iteration, so it was still high.
- Then, we took larger learning rate, but we saw that the loss did not converge to small value.
- At last, we took learning rate to balance the results from the two last tries, that is not too high and not too small, and we saw that the loss got smaller in the training process.

Learning curve- Loss per epoch

The train loss monotonically decrease between epochs.

The validation loss relatively follows the train loss, and their loss values are close to each other.

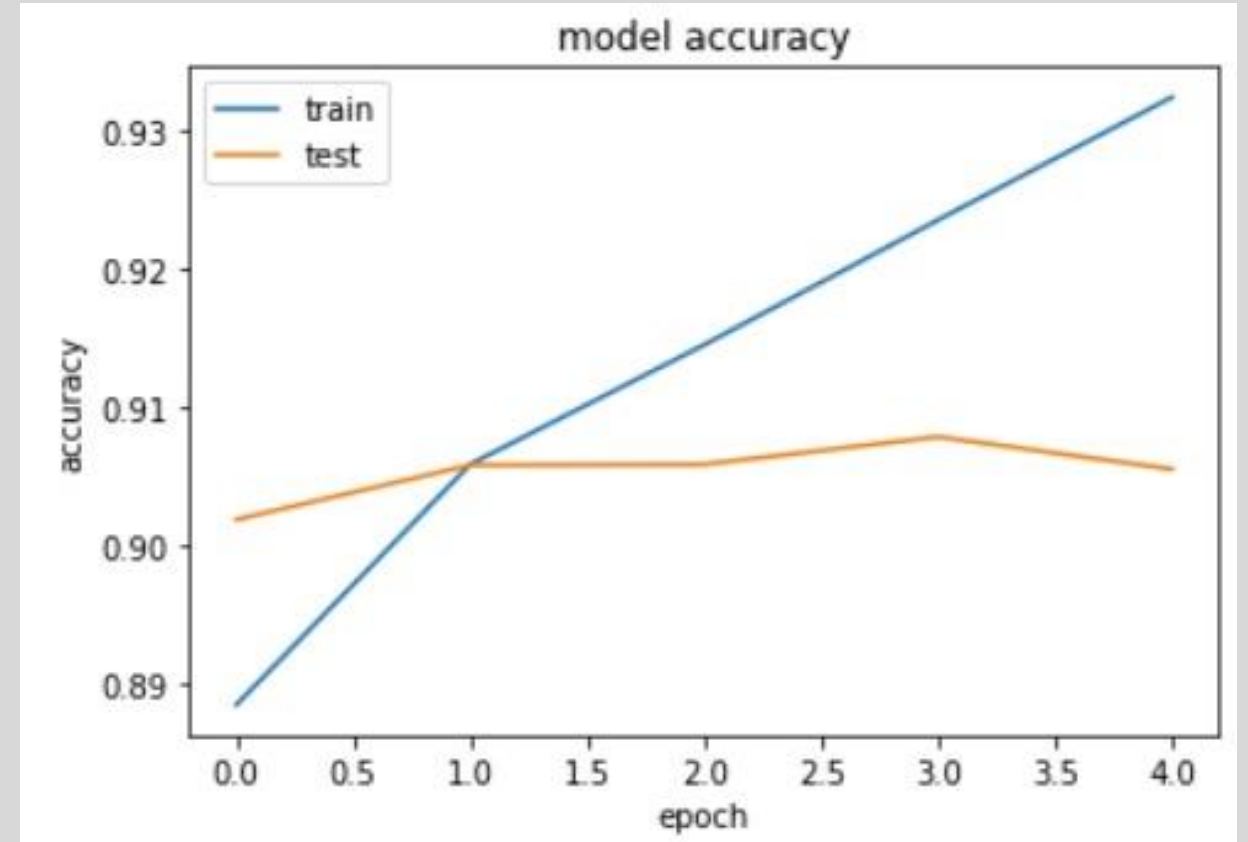


Learning curve- Accuracy per epoch

We can see that the validation curve followed the train curve with difference smaller than 0.03.

The accuracy of the validation is a little smaller than the train accuracy, as expected.

We can conclude that the trained model is not overfitting.



Predictions

- **F1 score:**

Lead 0.814042786615469

Position 0.673582295988935

Claim 0.6026842778014486

Evidence 0.6847481507573089

Concluding Statement 0.7957235586101565


Counterclaim 0.46105919003115264

Rebuttal 0.3746556473829201

Overall 0.6294994153124843

Our submission

- **Submission:**

Fork of feedback prize competition b2195b (version 6/8) 2 days ago by Noam Atias final feedback prize competition b2195b Version 6	Succeeded	0.620	0.607	
--	-----------	-------	-------	---

- **Public score: 0.607**
- **Private score: 0.620**
- **Place in competition:**
- **Private leaderboard: 1607**
- **Public leaderboard: 1606**